CLAIMS

What is claimed is:

1. In a computer-based system, a method of building a statistical model, comprising:

automatically identifying and flagging categorical variables in a data set containing both categorical and continuous variables;

automatically identifying categorical variables that are correlated with one or more continuous variables and eliminating categorical variable that are correlated with at least one continuous variable from a training data matrix used to build a statistical model, wherein the training data matrix comprises a subset of the original data set; and

building the statistical model based on the training data matrix.

2. The method of claim 1 wherein said step of automatically identifying and flagging categorical variables comprises:

determining if a variable contains integer observation values;

if the variable contains integer values, determining the number of unique integer values contained in the variable;

determining if the number of unique values exceeds a predetermined threshold value; and

if the number of unique values does not exceed the threshold value, flagging the variable as a categorical variable.

3. The method of claim 2 further comprising:

if the number of unique values exceeds the threshold value, determining if the variable has predictive strength greater than a predetermined value of Pearson's r;

if the variable has predictive strength greater than the predetermined value of Pearson's r, flagging the variable as a continuous variable;

if the variable has predictive strength less than the predetermined value of Pearson's r,

reducing the number of unique values by eliminating those unique values containing less than a predetermined number of entries so as to create a reduced variable set with a reduced number of unique values;

determining if the reduced number of unique values exceeds the threshold value; and

if the reduced number of unique values does not exceed the threshold value, flagging the variable as a categorical variable, else flagging the variable as a continuous variable.

4. The method of claim 1 wherein said step of automatically identifying categorical variables that are highly correlated with one or more continuous variables comprises:

binning at least one continuous variable so as to convert the continuous variable into a psuedo-categorical variable; and

calculating a Cramer's V value between at least one categorical variable and the psuedo-categorical variable to obtain an estimated measure of co-linearity between the categorical variable and the continuous variable.

5. The method of claim 1 further comprising:

calculating a correlation value for each variable in the training data matrix with respect to a target variable;

sorting the variables based on their correlation with the target variable; and

retaining a predetermined number of variables having the highest correlation values and eliminating any remaining variables from the training data matrix.

6. The method of claim 1 further comprising:

expanding each categorical variable contained in the training data matrix into a plurality of dummy variables;

measuring a predictive strength for each dummy variable and continuous variable in the training data matrix toward a target variable;

determining if any pair of variables in the set of dummy and continuous variables exhibits a pair-wise correlation greater than a predetermined threshold; and

if a pair of variables exhibits a pair-wise correlation greater than the threshold, eliminating one of the variables in the pair from the training data matrix, wherein the eliminated variable exhibits less predictive strength toward the target variable than the non-eliminated variable in the pair.

7. The method of claim 1 further comprising:

creating a plurality of principle components from the variables contained in the training data matrix, wherein each principle component comprises a linear combination of variables;

sorting the plurality of principle components by how much variance of the training data matrix each component captures;

selecting a subset of the plurality of principle components that captures a variance greater than a predetermined percentage of total variance; and

using the selected principle components to build the statistical model.

8. The method of claim 7 wherein said step of using the selected principle components to build the statistical model comprises:

performing a singular value decomposition (SVD) to generate a loading matrix; and

mapping coefficients calculated for the principle components back to corresponding variables of the training data matrix using the loading matrix.

9. The method of claim 1 further comprising:

performing a singular value decomposition (SVD) analysis using the variables contained in the training data matrix if the number of records in the training data matrix is less than a predetermined value; and

otherwise, performing a conjugate gradient descent (CGD) analysis on a residual sum

of squares based on the variables contained in the training data matrix if the number of records in the training data matrix is greater than or equal to the predetermined value.

10. The method of claim 1 further comprising:

detecting outlier values in the data set; and

for each detected outlier value, presenting a user with the following three options for handling the outlier value: (1) substitute the outlier value with a maximum or minimum non-outlier value in the data set; (2) keep the outlier value in the data set; (3) delete the record corresponding to the outlier value.

11. The method of claim 1 further comprising:

detecting missing values in the data set; and

for each missing value of a variable, inserting amean value of non-missing values of the variable in place of the missing value in the data set.

12. The method of claim 1 further comprising:

automatically detecting continuous variables having an exponential distribution; and

log-scaling those continuous variables using the following formula:

$$bx(i) = 1 - e^{-\frac{x(i)-\min}{mean-\min}},$$

where *x(i)* is a continuous variable being analyzed, *min, and mean* is the minimum value and the mean value of the variable in samples, respectively.

13. The method of claim 12 further comprising normalizing all the variables in the training data matrix.

14. The method of claim 1 further comprising randomly splitting the data set into a subset of training variables and a subset of test variables, wherein the training variables are used to

create the training data matrix for building the model and the subset of test variables are subsequently used to test the resulting model.

15. The method of claim 14 wherein prior to using the subset of test variables to test the model, pre-processing is performed on variables in the test set so as to create a test data matrix containing the same variables and same format as the training data matrix.

16. In a computer-based system, a method of building a statistical model, comprising:

automatically identifying and flagging categorical variables in a data set containing both categorical and continuous variables, wherein this step comprises:

determining if a variable contains integer observation values;

if the variable contains integer values, determining the number of unique integer values contained in the variable;

determining if the number of unique values exceeds a predetermined threshold value; and

if the number of unique values does not exceed the threshold value, flagging the variable as a categorical variable;

automatically identifying categorical variables that are correlated with one or more continuous variables and eliminating categorical variables that are correlated with at least one continuous variable from a training data matrix used to build a statistical model, wherein the training data matrix comprises a subset of the original data set; and

building the statistical model based on the training data matrix.

17. The method of claim 16 further comprising:

if the number of unique values exceeds the threshold value, determining if the variable has predictive strength greater than a predetermined value of Pearson's r;

if the variable has predictive strength greater than the predetermined value of Pearson's

r, flagging the variable as a continuous variable;

if the variable has predictive strength less than the predetermined value of Pearson's r, reducing the number of unique values by eliminating those unique values containing less than a predetermined number of entries so as to create a reduced variable set with a reduced number of unique values;

determining if the reduced number of unique values exceeds the threshold value; and

if the reduced number of unique values does not exceed the threshold value, flagging the variable as a categorical variable, else flagging the variable as a continuous variable.

18. The method of claim 16 further comprising:

creating a plurality of principle components from the variables contained in the training data matrix, wherein each principle component comprises a linear combination of variables;

sorting the plurality of principle components by how much variance of the training data matrix each component captures;

selecting a subset of the plurality of principle components that captures a variance greater than a predetermined percentage of total variance; and

using the selected principle components to build the statistical model.

19. The method of claim 18 wherein said step of using the selected principle components to build the statistical model comprises:

performing a singular value decomposition (SVD) to generate a loading matrix; and

mapping coefficients calculated for the principle components back to corresponding variables of the training data matrix using the loading matrix.

20. The method of claim 18 further comprising:

performing a singular value decomposition (SVD) analysis using the variables contained in the training data matrix if the number of records in the training data matrix is less

than a predetermined value; and

otherwise, performing a conjugate gradient descent (CGD) analysis on a residual sum of squares based on the variables contained in the training data matrix if the number of records in the training data matrix is greater than or equal to the predetermined value.

21. The method of claim 16 further comprising:

automatically detecting continuous variables having an exponential distribution; and

log-scaling those continuous variables using the following formula:

$$bx(i) = 1 - e^{-\frac{x(i)-\min}{mean-\min}},$$

where *x(i)* is a continuous variable being analyzed, *min, and mean* is the minimum value and the mean value of the variable in samples, respectively.

22. In a computer-based system, a method of building a statistical model, comprising:

automatically identifying and flagging categorical variables in a data set containing both categorical and continuous variables;

binning at least one continuous variable so as to convert the continuous variable into a psuedo-categorical variable;

calculating a Cramer's V value between at least one categorical variable and the psuedo-categorical variable to obtain an estimated measure of co-linearity between the categorical variable and the continuous variable;

based on the calculated Cramer's V value, eliminating a corresponding categorical variable that is correlated with at least one continuous variable from a training data matrix used to build a statistical model, wherein the training data matrix comprises a subset of the original data set; and

building the statistical model based on the training data matrix.

23. The method of claim 22 further comprising:

calculating a correlation value for each variable in the training data matrix with respect to a target variable;

sorting the variables based on their correlation with the target variable; and

retaining a predetermined number of variables having the highest correlation values and eliminating any remaining variables from the training data matrix.

24. The method of claim 22 further comprising:

expanding each categorical variable contained in the training data matrix into a plurality of dummy variables;

measuring a predictive strength for each dummy variable and continuous variable in the training data matrix toward a target variable;

determining if any pair of variables in the set of dummy and continuous variables exhibits a pair-wise correlation greater than a predetermined threshold; and

if a pair of variables exhibits a pair-wise correlation greater than the threshold, eliminating one of the variables in the pair from the training data matrix, wherein the eliminated variable exhibits less predictive strength toward the target variable than the non-eliminated variable in the pair.

25. The method of claim 22 further comprising:

creating a plurality of principle components from the variables contained in the training data matrix, wherein each principle component comprises a linear combination of two or more variables;

sorting the plurality of principle components by how much variance of the training data matrix each component captures;

selecting a subset of the plurality of principle components that captures a variance greater than a predetermined percentage of total variance; and

using the selected principle components to build the statistical model.

26. The method of claim 25 wherein said step of using the selected principle components to build the statistical model comprises:

performing a singular value decomposition (SVD) to generate a loading matrix; and

mapping coefficients calculated for the principle components back to corresponding variables of the training data matrix using the loading matrix.

27. The method of claim 25 further comprising:

performing a singular value decomposition (SVD) analysis using the variables contained in the training data matrix if the number of records in the training data matrix is less than a predetermined value; and

otherwise, performing a conjugate gradient descent (CGD) analysis on a residual sum of squares based on the variables contained in the training data matrix if the number of records in the training data matrix is greater than or equal to the predetermined value.

28. The method of claim 22 further comprising:

automatically detecting continuous variables having an exponential distribution; and

log-scaling those continuous variables using the following formula:

$$bx(i) = 1 - e^{-\frac{x(i)-\min}{mean-\min}},$$

where *x(i)* is a continuous variable being analyzed, *min, and mean* is the minimum value and the mean value of the variable in samples, respectively.

29. A computer-readable medium containing code executable by a computer that when executed performs a process of automatically building a statistical model, said process comprising:

automatically identifying and flagging categorical variables in a data set containing

both categorical and continuous variables;

automatically identifying categorical variables that are correlated with one or more continuous variables and eliminating categorical variables that are correlated with at least one continuous variable from a training data matrix used to build a statistical model, wherein the training data matrix comprises a subset of the original data set; and

building the statistical model based on the training data matrix.

30. The computer-readable medium of claim 29 wherein said step of automatically identifying and flagging categorical variables comprises:

determining if a variable contains integer observation values;

if the variable contains integer values, determining the number of unique integer values contained in the variable;

determining if the number of unique values exceeds a predetermined threshold value; and

if the number of unique values does not exceed the threshold value, flagging the variable as a categorical variable.

31. The computer-readable medium of claim 30 wherein said process further comprises:

if the number of unique values exceeds the threshold value, determining if the variable has predictive strength greater than a predetermined value of Pearson's r;

if the variable has predictive strength greater than the predetermined value of Pearson's r, flagging the variable as a continuous variable;

if the variable has predictive strength less than the predetermined value of Pearson's r, reducing the number of unique values by eliminating those unique values containing less than a predetermined number of entries so as to create a reduced variable set with a reduced number of unique values;

determining if the reduced number of unique values exceeds the threshold value; and

if the reduced number of unique values does not exceed the threshold value, flagging the variable as a categorical variable, else flagging the variable as a continuous variable.

32. The computer-readable medium of claim 29 wherein said step of automatically identifying categorical variables that are highly correlated with one or more continuous variables comprises:

binning at least one continuous variable so as to convert the continuous variable into a psuedo-categorical variable; and

calculating a Cramer's V value between at least one categorical variable and the psuedo-categorical variable to obtain an estimated measure of co-linearity between the categorical variable and the continuous variable.

33. The computer-readable medium of claim 29 wherein said process further comprises:

calculating a correlation value for each variable in the training data matrix with respect to a target variable;

sorting the variables based on their correlation with the target variable; and

retaining a predetermined number of variables having the highest correlation values and eliminating any remaining variables from the training data matrix.

34. The computer-readable medium of claim 29 wherein said process further comprises:

expanding each categorical variable contained in the training data matrix into a plurality of dummy variables;

measuring a predictive strength for each dummy variable and continuous variable in the training data matrix toward a target variable;

determining if any pair of variables in the set of dummy and continuous variables exhibits a pair-wise correlation greater than a predetermined threshold; and

if a pair of variables exhibits a pair-wise correlation greater than the threshold,

eliminating one of the variables in the pair from the training data matrix, wherein the eliminated variable exhibits less predictive strength toward the target variable than the non-eliminated variable in the pair.

35. The computer-readable medium of claim 29 wherein said process further comprises:

creating a plurality of principle components from the variables contained in the training data matrix, wherein each principle component comprises a linear combination of variables;

sorting the plurality of principle components by how much variance of the training data matrix each component captures;

selecting a subset of the plurality of principle components that captures a variance greater than a predetermined percentage of total variance; and

using the selected principle components to build the statistical model.

36. The computer-readable medium of claim 35 wherein said step of using the selected principle components to build the statistical model comprises:

performing a singular value decomposition (SVD) to generate a loading matrix; and

mapping coefficients calculated for the principle components back to corresponding variables of the training data matrix using the loading matrix.

37. The computer-readable medium of claim 35 wherein said process further comprises:

performing a singular value decomposition (SVD) analysis using the variables contained in the training data matrix if the number of records in the training data matrix is less than a predetermined value; and

otherwise, performing a conjugate gradient descent (CGD) analysis on a residual sum of squares based on the variables contained in the training data matrix if the number of records in the training data matrix is greater than or equal to the predetermined value.

38. The computer-readable medium of claim 29 wherein said process further comprises:

detecting outlier values in the data set; and

for each detected outlier value, presenting a user with the following three options for handling the outlier value: (1) substitute the outlier value with a maximum or minimum non-outlier value in the data set; (2) keep the outlier value in the data set; (3) delete the record corresponding to the outlier value.

39. The computer-readable medium of claim 29 wherein said process further comprises:

detecting missing values in the data set; and

for each missing value of a variable, inserting amean value of non-missing values of the variable in place of the missing value in the data set.

40. The computer-readable medium of claim 29 wherein said process further comprises:

automatically detecting continuous variables having an exponential distribution; and

log-scaling those continuous variables using the following formula:

$$bx(i) = 1 - e^{-\frac{x(i)-\min}{mean-\min}},$$

where *x(i)* is a continuous variable being analyzed, *min, and mean* is the minimum value and the mean value of the variable in samples, respectively.

41. The computer-readable medium of claim 40 wherein said process further comprises normalizing all the variables in the training data matrix.

42. The computer-readable medium of claim 29 wherein said process further comprises randomly splitting the data set into a subset of training variables and a subset of test variables, wherein the training variables are used to create the training data matrix for building the model and the subset of test variables are subsequently used to test the resulting model.

43. The computer-readable medium of claim 42 wherein prior to using the subset of test variables to test the model, pre-processing is performed on variables in the test set so as to create a test data matrix containing the same variables and same format as the training data matrix.

44. A computer-readable medium containing code executable by a computer that when executed performs a process of automatically building a statistical model, the process comprising:

automatically identifying and flagging categorical variables in a data set containing both categorical and continuous variables, wherein this step comprises:

determining if a variable contains integer observation values;

if the variable contains integer values, determining the number of unique integer values contained in the variable;

determining if the number of unique values exceeds a predetermined threshold value; and

if the number of unique values does not exceed the threshold value, flagging the variable as a categorical variable;

automatically identifying categorical variables that are correlated with one or more continuous variables and eliminating categorical variables that are correlated with at least one continuous variable from a training data matrix used to build a statistical model, wherein the training data matrix comprises a subset of the original data set; and

building the statistical model based on the training data matrix.

45. The computer-readable medium of claim 44 wherein said process further comprises:

if the number of unique values exceeds the threshold value, determining if the variable has predictive strength greater than a predetermined value of Pearson's r;

if the variable has predictive strength greater than the predetermined value of Pearson's r, flagging the variable as a continuous variable;

if the variable has predictive strength less than the predetermined value of Pearson's r, reducing the number of unique values by eliminating those unique values containing less than a predetermined number of entries so as to create a reduced variable set with a reduced number of unique values;

determining if the reduced number of unique values exceeds the threshold value; and

if the reduced number of unique values does not exceed the threshold value, flagging the variable as a categorical variable, else flagging the variable as a continuous variable.

46. The computer-readable medium of claim 44 wherein said process further comprises:

creating a plurality of principle components from the variables contained in the training data matrix, wherein each principle component comprises a linear combination of variables;

sorting the plurality of principle components by how much variance of the training data matrix each component captures;

selecting a subset of the plurality of principle components that captures a variance greater than a predetermined percentage of total variance; and

using the selected principle components to build the statistical model.

47. The computer-readable medium of claim 46 wherein said step of using the selected principle components to build the statistical model comprises:

performing a singular value decomposition (SVD) to generate a loading matrix; and

mapping coefficients calculated for the principle components back to corresponding variables of the training data matrix using the loading matrix.

48. The computer-readable medium of claim 46 wherein said process further comprises:

performing a singular value decomposition (SVD) analysis using the variables contained in the training data matrix if the number of records in the training data matrix is less than a predetermined value; and

otherwise, performing a conjugate gradient descent (CGD) analysis on a residual sum of squares based on the variables contained in the training data matrix if the number of records in the training data matrix is greater than or equal to the predetermined value.

49. The computer-readable medium of claim 46 wherein said process further comprises:

automatically detecting continuous variables having an exponential distribution; and

log-scaling those continuous variables using the following formula:

$$bx(i) = 1 - e^{-\frac{x(i)-\min}{mean-\min}},$$

where *x(i)* is a continuous variable being analyzed, *min, and mean* is the minimum value and the mean value of the variable in samples, respectively.

50. A computer-readable medium containing code executable by a computer that when executed performs a process of automatically building a statistical model, the process comprising:

automatically identifying and flagging categorical variables in a data set containing both categorical and continuous variables;

binning at least one continuous variable so as to convert the continuous variable into a psuedo-categorical variable;

calculating a Cramer's V value between at least one categorical variable and the psuedo-categorical variable to obtain an estimated measure of co-linearity between the categorical variable and the continuous variable;

based on the calculated Cramer's V value, eliminating a corresponding categorical variable that is correlated with at least one continuous variable from a training data matrix used to build a statistical model, wherein the training data matrix comprises a subset of the original data set; and

building the statistical model based on the training data matrix.

51. The computer-readable medium of claim 50 wherein said process further comprises:

calculating a correlation value for each variable in the training data matrix with respect to a target variable;

sorting the variables based on their correlation with the target variable; and

retaining a predetermined number of variables having the highest correlation values and eliminating any remaining variables from the training data matrix.

52. The computer-readable medium of claim 50 wherein said process further comprises:

expanding each categorical variable contained in the training data matrix into a plurality of dummy variables;

measuring a predictive strength for each dummy variable and continuous variable in the training data matrix toward a target variable;

determining if any pair of variables in the set of dummy and continuous variables exhibits a pair-wise correlation greater than a predetermined threshold; and

if a pair of variables exhibits a pair-wise correlation greater than the threshold, eliminating one of the variables in the pair from the training data matrix, wherein the eliminated variable exhibits less predictive strength toward the target variable than the non-eliminated variable in the pair.

53. The computer-readable medium of claim 50 wherein said process further comprises:

creating a plurality of principle components from the variables contained in the training data matrix, wherein each principle component comprises a linear combination of variables;

sorting the plurality of principle components by how much variance of the training data matrix each component captures;

selecting a subset of the plurality of principle components that captures a variance greater than a predetermined percentage of total variance; and

using the selected principle components to build the statistical model.

54. The computer-readable medium of claim 53 wherein said step of using the selected principle components to build the statistical model comprises:

performing a singular value decomposition (SVD) to generate a loading matrix; and

mapping coefficients calculated for the principle components back to corresponding variables of the training data matrix using the loading matrix.

55. The computer-readable medium of claim 53 further comprising:

performing a singular value decomposition (SVD) analysis using the variables contained in the training data matrix if the number of records in the training data matrix is less than a predetermined value; and

otherwise, performing a conjugate gradient descent (CGD) analysis on a residual sum of squares based on the variables contained in the training data matrix if the number of records in the training data matrix is greater than or equal to the predetermined value.

56. The computer-readable medium of claim 50 further comprising:

automatically detecting continuous variables having an exponential distribution; and

log-scaling those continuous variables using the following formula:

$$bx(i) = 1 - e^{-\frac{x(i)-\min}{mean-\min}},$$

where *x(i)* is a continuous variable being analyzed, *min, and mean* is the minimum value and the mean value of the variable in samples, respectively.